Operations Research Letters 44 (2016) 796–800

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Heavy-traffic limits for a single-server queue leading up to a critical point

Ward Whitt

Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027-6699, USA

ARTICLE INFO

Article history:
Received 10 August 2016
Received in revised form
21 September 2016
Accepted 6 October 2016
Available online 14 October 2016

Keywords:
Heavy-traffic limits
Nonstationary queues
Queues with time-varying arrival rates
Onset of critical loading
Rush hour
Transition through saturation

ABSTRACT

We establish heavy-traffic limits for the arrival and workload processes in a single-server queue with a time-varying arrival-rate function. We establish limits at and before a critical point, the onset of critical loading, where the arrival-rate function approaches its critical value from below. We extend results by Newell (1968) and Mandelbaum and Massey (1995) and present alternative views of the interesting scaling constants that arise in these limits.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

The purpose of this paper is to extend, and contribute to a better understanding of, diffusion approximations and heavy-traffic limits for a single-server queue with time-varying arrival-rate function that were established by Newell [15–18] and Mandelbaum and Massey [11]; also see [12,8,13,4]. As in our recent paper [22] for queues with periodic arrival-rate functions, we develop these limits in the standard framework for heavy-traffic limits, as in [6,7,19,21].

In this paper we focus on one special case: the limiting behavior at and before an isolated critical point, at what is called the onset of critical loading in [11], where the arrival rate approaches the critical value from below. Thus, this paper relates to only Theorem 3.4 in [11]. For that result, we generalize the setting from $M_t/M/1$ to $G_t/G/1$ and give alternative developments, leading to alternative scaling and alternative interpretations of it.

In particular, we consider a single-server queue with unlimited waiting room having service times with mean 1, which fixes the time scale. We consider a sequence of models with an associated sequence of arrival processes having time-varying arrival-rate functions. We will establish heavy-traffic limits, which involve scaling time, so that we are looking at intervals over which many customers arrive and are served. We assume that there is an isolated critical point, which we take to be time 0. In particular, we assume that the arrival-rate function satisfies

$$\lambda(0) = 1 \quad \text{and} \quad \lambda(t) < 1 \quad \text{for } t < 0. \tag{1}$$

Moreover, for simplicity, we assume that λ is nondecreasing in t before time 0. (There is no mass of workload from the distant past contributing to the buildup of congestion at the critical point.) As observed in [15], the congestion at times $t < 0$ is less, often much less, than the steady-state distribution with the instantaneous traffic intensity $\rho(t) = \lambda(t)$, because the traffic intensity was previously at lower values.

The approaches in [15–18,11] are quite different from [6,7,19,21], even though they can be related. First, Newell [15–18] makes a direct diffusion process approximation and then analyzes the Fokker–Planck partial differential equation for the time-varying cumulative distribution function. The papers [15–17] are landmark contributions to queueing theory, but they are challenging to understand, because they both develop the diffusion approximation and analyze it. It turns out that the diffusion process is the limiting diffusion in the heavy-traffic limits, with the key asymptotic properties captured by the parameters of that diffusion process. In contrast, as in [6,7,19,21], our approach emphasizes scaling, so it avoids looking directly at the detailed evolution of the diffusion process, but that remains to be done to calculate explicit approximations. In [23] we develop a robust queueing approach to calculate such explicit approximations.

The more modern [11] exploits strong approximations. For scaling, it starts with an initial arrival-rate function and expands

E-mail address: ww2040@columbia.edu.

http://dx.doi.org/10.1016/j.orl.2016.10.005
0167-6377/© 2016 Elsevier B.V. All rights reserved.

time about each fixed point in that arrival function. As a consequence of that expansion, the relevant long-time behavior of the arrival-rate function is determined by the local behavior of the initial function, as exposed by a Taylor-series expansion, which requires extra regularity assumptions. This approach is helpful for analyzing highly structural mathematical models, as we illustrate with a sinusoidal example in Example 4.1.

Here is how this paper is organized. In Section 2 we formulate our arrival process model. In Section 3 we establish a heavy-traffic functional central limit theorem (FCLT) for the arrival process. In Section 4 we show how that arrival process FCLT can be recast in a setting in which we expand time within a fixed initial arrival-rate function, as in [11].

Motivated by the desire to develop an approach that is helpful for applications, in Section 5 we show how the heavy-traffic FCLT can be expressed in yet a different way using drift scaling, which is in the spirit of §4.3 of [20]. We think that it is natural to first fit the drift function and then afterwards choose the appropriate time and space scaling that goes with that drift function. That approach directly yields the appropriate space scaling and the diffusion process approximation, which the FCLT's imply should perform well when the drift constant is suitably small. Given a FCLT for the arrival process, corresponding heavy-traffic FCLT's follow for the standard queueing processes using the continuous mapping approach in [21]. In Section 6 we illustrate by establishing the heavy-traffic FCLT for the workload process. Finally, we briefly draw conclusions in Section 7.

2. The arrival process model

As in [22], we construct the arrival process A by composing a process assumed to satisfy a FCLT and a deterministic cumulative arrival-rate function. In particular, we let the stochastic arrival counting processes defined by

$$A(t) \equiv N(\Lambda(t)), \quad t \geq 0, \tag{2}$$

where N is a stochastic counting process satisfying a FCLT, i.e.,

$$\hat{N}_n(t) \equiv n^{-1/2}[N(nt) - nt] \Rightarrow c_a B_a(t) \quad \text{in } \mathcal{D} \text{ as } n \to \infty, \tag{3}$$

where $\Rightarrow$ denotes convergence in distribution in the function space $\mathcal{D}$ of right-continuous real-valued functions on the interval $[0, \infty)$ with left limits, as in [21], and B_a is a standard (drift 0, variance 1) Brownian motion (BM), while Λ is a cumulative arrival-rate function, satisfying $\Lambda(t) \equiv \int_0^t \lambda(s)\, ds$, $t \geq 0$, with λ being the arrival-rate function, which is assumed to be integrable over finite intervals.

The construction in (2) is convenient for constructing non-Markov nonstationary arrival processes. It was suggested in [14] and also used in [3,5,22,23]. However, it is important to recognize that, even though it allows very general stochastic processes N, including renewal processes and much more (see §4.4 of [21]), this model is highly structured, having all unpredictable stochastic variability associated with the process N, with its FCLT behavior captured by the single variability parameter c_a, while all the predictable deterministic variability associated with the deterministic arrival-rate function λ and its associated cumulative rate function Λ. More generally, we might contemplate a time-varying variability parameter. In the present context, if the process N is a renewal counting process, then c_a is the square root of c_a^2, the squared coefficient of variation (scv, variance divided by the square of the mean) of an interarrival time. From an engineering perspective, the tractability produced by reducing the impact of the stochastic variability to the single parameter c_a^2 may be essential for drawing useful conclusions about system performance.

Throughout this paper, we assume that the cumulative arrival-rate function Λ is deterministic, but it is significant that the results here can be extended to cover the case in which arrival-rate function is a stochastic process, which can be important in applications. For example, service system arrival process data often indicate overdispersion caused by day-to-day variation, as discussed in [9].

3. A conventional heavy-traffic FCLT for the arrival process

Given the composition representation of the arrival process in (2) and the assumed FCLT in (3), we can obtain a conventional FCLT for the arrival process A defined in (2), which involves scaling time by n and space by $1/\sqrt{n}$, and then letting $n \to \infty$, if we write

$$\hat{A}_n(t) \equiv n^{-1/2}[A_n(nt) - nt] \quad \text{and}$$
$$\hat{\Lambda}_n(t) \equiv n^{-1/2}[\Lambda_n(nt) - nt], \quad t \geq 0, \tag{4}$$

where

$$A_n(t) \equiv N(\Lambda_n(t)) \quad \text{and} \quad \Lambda_n(t) \equiv \int_0^t \lambda_n(s)\, ds, \quad t \geq 0, \tag{5}$$

and we make appropriate assumptions about the deterministic arrival-rate functions $\lambda_n(t)$, which requires that $\lambda_n(t)$ remain close to 1 for large time intervals about $t = 0$. As in [22], it is important that we scale time and space in the deterministic cumulative arrival rate functions Λ_n in (4).

We find it convenient to work in reverse time, because the workload process then can be represented as a simple supremum of the net input process; see Section 6. The reverse-time construction is discussed in [23], which develops a time-varying robust queueing approximation based on the supremum representation. Hence, we measure time backwards from time 0, so that $A(t)$ counts the number of arrivals in $[-t, 0]$. This section is devoted to establishing a FCLT for the process A. We remark that a FCLT also holds in forward time or in intervals $[t_1, t_2]$ with $t_1 < 0 < t_2$ by the same argument.

As a main example in our reverse-time framework, we focus on arrival-rate functions that decay in a power away from the critical point at time 0; i.e.,

$$\lambda_n(t) = 1 - c_n t^p, \quad t \geq 0, \tag{6}$$

for some real number $p \geq 0$. In comparison to [11] and Section 4, note that p is not restricted to being an integer, but $p = 1$ and $p = 2$ are especially interesting.

We emphasize that we are concerned with large time, i.e., time scaled by n as $n \to \infty$, so that we need to consider the scaled version

$$\lambda_n(nt) = 1 - c_n(nt)^p, \quad t \geq 0, \tag{7}$$

where we allow $n \to \infty$. Note that $p = 0$ in (6) and (7) corresponds to a constant arrival rate of $1 - c_n$, which produces a constant negative drift of $1 - c_n$. The following result covers this stationary model as a special case.

Theorem 3.1 (*Conventional FCLT for the Arrival Process*). *If, in addition to the FCLT for $\hat{N}_n$ in (3),*

$$\hat{\Lambda}_n \to \hat{\Lambda} \quad \text{in } \mathcal{D}, \tag{8}$$

for $\hat{\Lambda}_n$ defined in (4) and (5), then

$$\hat{A}_n \Rightarrow c_a B_a + \hat{\Lambda} \quad \text{in } \mathcal{D} \text{ as } n \to \infty. \tag{9}$$

Under assumption (6), the limit in (8) holds if and only if

$$c_n n^{(2p+1)/2} \to c \quad \text{as } n \to \infty, \ 0 < c < \infty, \tag{10}$$

Download English Version:

https://daneshyari.com/en/article/5128415

Download Persian Version:

https://daneshyari.com/article/5128415

Daneshyari.com